

Using Visualization to Explore Original and Anonymized LBSN Data

EuroVis 2016, University of Calgary

背景及动机

- 对象
 - LBSN, location-based social network。
 - 数据隐私
- 术语
 - 用户: 社交网络中的用户
 - 分析者: 分析社交网络数据的人
 - 地点 (CI, check-in): LBSN中用户访问过的一个地方
- 问题: LBSN数据通常带有用户敏感信息, 因此对于数据分析者来说, 对社交网络用户进行匿名化有助于降低信息泄漏。
- 与隐私问题专家进行合作。

贡献

- GSUVis系统: 更好地在保障隐私的前提下对LBSN数据进行探索。
- 在数据隐私环境下的数据探索与分析: 列出了一些任务类型, 这些任务也可能用于活动轨迹、电话记录或商品购买记录的探索。
- Evaluation: 新的发现和对比实验。

概览

- 数据集: Gowalla (公开数据集)
- Tasks: 下图

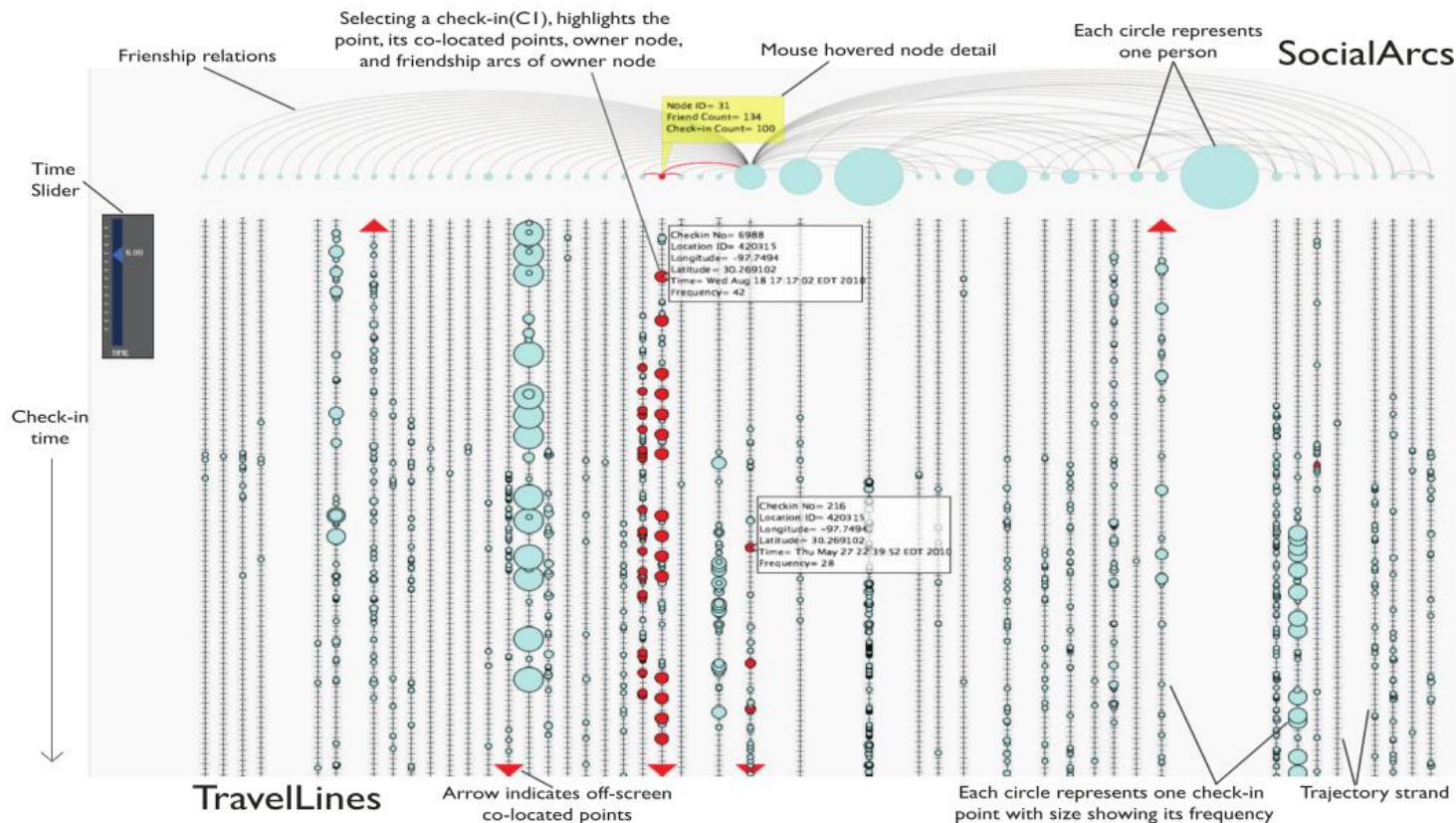
Tasks for exploring the original LBSN dataset		Data	P1	P2	P3	P4	P5	P6
T1	Identify the popularity of members based on the number of social connections.							
T2	Identify the co-located check-in points.							
T3	Identify the co-located check-in points occuring at about the same time.							
T4	Identify a person's favorite locations based on the number of check-ins within her location trajectory.							
T5	What is the most/least popular location among members?							
T6	Is there any correlation between social connections and check-in points?							
T7	Is there any pattern in terms of the check-in behaviour of an individual?							
T8	Compare the check-in behaviour of two (or more) friends.							
T9	What is the trajectory of an individual based on location type? (e.g., home, hospital, home, shopping mall)							
T10	Order check-in points according to their occurrence time.							
T11	What is the temporal frequency of location check-ins for an individual?							
T12	Are there any outliers in the sequence of declared locations of a person?							
Tasks for exploring the anonymized LBSN dataset								
T13	Identify the data loss in the social graph (e.g., amount of adeleted edges).							
T14	Identify the data loss in the location trajectory of each individual (e.g., movement of anonymized point vs original ones).							
T15	For which cases the data utility was/wasn't well preserved?							
T16	Compare different (two or more) anonymization algorithms in terms of data utility.							
T17	Is there a correlation between more/less social people and their data utility?							
T18	Is there a correlation between people who have more/less check-in points and preserving their data utility?							
T19	Which location types (ie.g., hospital, shopping center) preserved after anonymization?							
T20	Is there any outliers for co-located data after anonymization? (e.g., co-located points mapped to different locations)							

image1

- Design goals
 - 布局紧凑: 需要同时展示社交关系和地理信息
 - 大规模network下局部位置的可读性
 - 可调整的视觉映射: LoD
 - 全局视图
 - 匿名化过程中整个视图内容的变化可见

设计

整体界面



两大部分：SocialArcs和TravelLines

SocialArcs

- 设计：横向排列的node-link（node-arcs）视图，节点代表社交网络中的用户，边（arcs）代表用户间的联系。
- 交互：鼠标选中节点，连接的arcs和对应的TravelLine红色高亮。

TravelLines

- 设计：每个用户节点下方引出竖线，自上至下代表时间从远到近。竖线上的圆圈代表用户在该时间点访问过某地点，圆圈大小代表。左侧有时间窗调整部件。
- 交互：选择一列中的一个圈（即一个地点），上方SocialArcs中对应的node同时会被选中。整个视图范围内代表相同地点的所有圈都会红色高亮（其他用户可能也访问过该选中的地点）。当访问该选中地点的时间点在当前时间窗范围外，则用竖线上方或下方的箭头表示。

保护隐私的SocialArcs和TravelLines

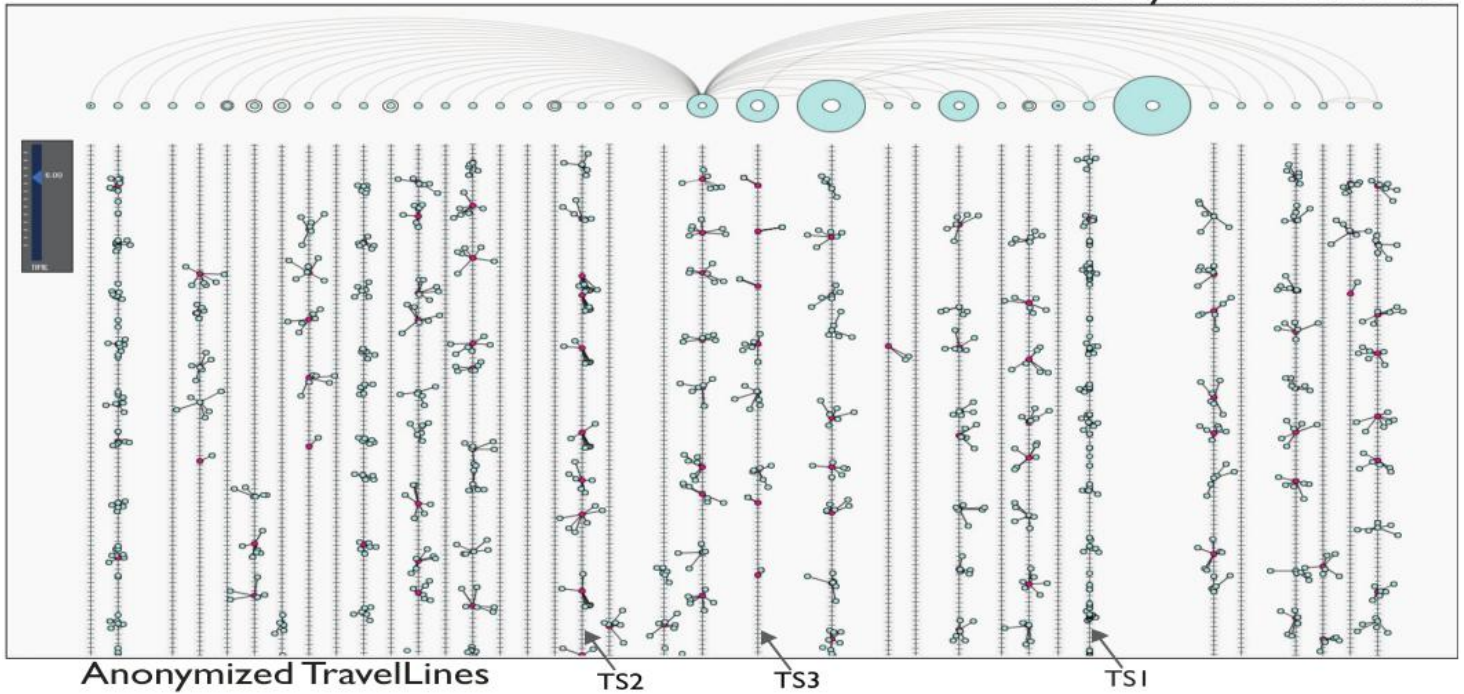


image1

- 使用一些graph anonymization方法将社交网络数据处理一遍，会随机添加、删除一些边。
- 对于SocialArcs中的节点，会在原来的蓝色圆上叠加一个白色同心圆，用于表示节点的度数。白色圆半径大于原来的蓝色圆则代表度数在匿名化后有增加；反之亦然。
- TravelLines中，地点被匿名化，圆圈大小不再编码信息。check in位置由一组位置的方向组合表达，用star glyph编码，仅表达出用户check in的位置在某些位置的那个方向。

实验

- 形式：专家review。
- 结论：专家一致好评。

数据探索及知识发现

- 三方面：用户行为发现；假设产生及验证；匿名化后的数据分析评估。
- 专家：匿名化后也能发现一些用户的行为模式。

专家意见

- 希望支持不同匿名算法的横向比较，以探索不同匿名算法对用户行为模式的匿名化和模式保持的能力。
- 希望能保持一些公共地点信息，例如车站、地标性建筑等。

未来工作

- Scalability
- 更多的数据类型

感想

- 数据隐私保护可能将成为一个热点。
- 方法部分的叙述采用虚拟分析者的形式（文中的Emma）。有利于增强代入感和描述交互操作。
- 该文章对visual design的描述极为详细，并写明每个设计和每个交互是为了完成某个task或是遵从某个design goal。